NOT TO BE FURTHER COPIED OR REPRODUCED

(* BATCHES *)

1

Scratching Down Numbers

(stem-and-leaf)

		Exh.		Page
	General Comment 1-0		a	
	- index pages 1-0		b	
	- exhibits 1-0	1	c	1-7
	- index entries 1-0	2	d	1-9
	- use of dashes 1-0	3		1-11
1A.	Quantitative detective work 1-1	4	e	1-15
1B.	Practical arithmetic 1-4	5		1-16
	- precision of arithmetic 1-4	6		1-18
	- rounding 1-4 to 1-5	7		1-20
	- growing 1-5	8	f	1-23/2-
	- units and decimal points 1-5a to 1-5b	9		1-25
	- h for "and a half" 1-5b	10	g	1-28
	- # for count 1-5c		h	
1C.	Scratching down numbers 1-5c	11	k	1-33
	batch defined at 1-5c	12	p	1-35
1D.	Doing better with stem-and-leaf 1-8	13		1-36
	stem-and-leaf displays, at 1-10	14		1-37
	- checking 1-12 to 1-14			
1E.	Using the right number of stems 1-14			
	- expanded stem-and-leaf 1-14			
	- compressed stem-and-leaf 1-14 to 1-16			
	- using mixed leaves 1-17 to 1-20			
1F.	How to count by tallying 1-21			
	- used cars again 1-22 to 1-26			
1G.	What does it mean to feel "what the data are like"? 1-26			
	**********			
1H.	How far have we come? 1-29			
	**********			
1K.	How to use stem-and-leaf to pick up additional information (optional technique) 1-32			
	**********			
1P.	Problems 1-34			

Comments About the Preceding Index Page

The index page for this chapter -- and for others -- is nonstandard. At least two aspects deserve note.

* exhibits *

To the right of the index page we give pages for numbered "exhibits". This neutral word is used for tables, charts, graphs, plots and the like. They are all numbered in a single sequence within each chapter, so they will be easy to find.

They will be referred to in text as "exhibit 14" without additional capitalization. Their first main appearance will be identified by using bold numbers, as **14**, for that appearance and not for others.

* index entries *

Two sorts of additional entries are given between the section titles:

- subtopics, entered with an initial "-" in the index, and centered in text between *'s.
- vital terms defined, marked as such in the index when short, and displayed like formulas in text.

* usage of dashes *

The comments just made also illustrate some textual usages that recur very frequently below:

- the use of ":" and hanging entries beginning with "-" to make lists obvious.
- the use of "--" to enclose parenthetic entries that can be thought of as possibly spoken by a different voice.

1

Scratching Down Numbers

(stem-and-leaf)

1A. Quantitative detective work

Exploratory data analysis is detective work -- numerical detective work -- or counting detective work -- or graphical detective work.

A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places. Equally, the analyst of data needs both tools and understanding. It is the purpose of this introduction to provide some of each.

Time will keep us from learning about many tools -- we shall try to look at a few of the most general and powerful among the simple ones. We do not guarantee to introduce you to the "best" tools, particularly since we are not sure that there can be unique bests.

Understanding has different limitations. As many detective stories have made clear, one needs quite different sorts of detailed understanding to detect criminals in London's slums, in a remote Welsh village, among Parisian aristocrats, in the cattle-raising west, or in the Australian outback. We do not expect a Scotland Yard officer to do well trailing cattle thieves, or a Texas ranger to be effective in the heart of Birmingham. Equally, very different detailed understandings are needed to be highly effective in dealing

with data concerning earthquakes, data concerning techniques of chemical manufacturing, data concerning the sizes and profits of firms in a service industry, data concerning human hearing, data concerning suicide rates, data concerning population growth, data concerning fossil dinosaurs, data concerning the genetics of fruit flies, or data concerning the latest exploits in molecular biology. A full introduction to data analysis in any one of these fields -- or to any of many others -- would take much more time than our whole introduction.

The Scotland Yard detective, however, would be far from useless in the wild west or the outback. He has certain general understandings of conventional detective work that will help him anywhere.

In data analysis there are similar general understandings. We can hope to lead you to a few of them. We shall try.

The processes of criminal justice are clearly divided between the search for the evidence -- in Anglo-Saxon hands the responsibility of the police and other investigative forces -- and the evaluation of the evidence's strength -- finally a matter for juries and judges. It is helpful to make a similar distinction in data analysis. Exploratory data analysis is detective in character. Confirmatory data analysis is judicial or quasi-judicial in character. Only exploratory data analysis will be our subject here.

Unless the detective finds the clues, judge or jury has nothing to consider. Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider.

Experiments and certain planned inquiries provide some exceptions and partial exceptions to this rule. They do this because one line of data analysis was planned as part of the experiment or inquiry. Even here, however, restricting one's self to the planned analysis -- failing to accompany it with exploration -- loses sight of the most interesting results too frequently to be comfortable.

As all detective stories remind us, many of the circumstances surrounding a crime are accidental or misleading. Equally, many of the indications to be discerned in bodies of data are accidental or misleading. To accept all appearances as conclusive would be destructively foolish, either in crime detection or in data analysis. To fail to collect all appearances because some -- or even most -- are only accidents would, however, be gross misfeasance deserving (and often receiving) appropriate punishment.

Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone -- as the first step.

We will be exploring numbers. We need to handle them easily and look at them effectively. Techniques for handling

and looking -- whether graphical, arithmetic, or intermediate -- will be important. The simpler we can make these techniques the better -- so long as they work, and work well. When details make an important difference, they deserve -- and will get -- emphasis.

1B. Practical arithmetic

A few arithmetic details deserve our attention before we go further.

* precision of arithmetic *

We shall feel free to be inconsistent about keeping extra decimal places or rounding off. We shall try to use reasonable care and good judgment in view of the total behavior of the data involved. Any or all of the following may occur:

$\frac{1}{3}(19+19+20) = 19.3$ (rest of data close by)

$\frac{1}{3}(19+19+20) = 19$ (rest of data farther away)

* rounding *

When we do round, it will be to the nearest number, with ties broken by preferring even numbers. Thus:

$$17.5 \to 18$$
$$18.5 \to 18$$
$$19.5 \to 20$$
$$20.5 \to 20$$

As a result, we may have

$\frac{1}{2}(19+20) = 19.5$ (rest of data close by)

$\frac{1}{2}(19+20) = 20$ (rest of data farther away)

$\frac{1}{2}(20+21) = 20.5$ (rest of data close by)

$\frac{1}{2}(20+21) = 20$ (rest of data farther away)

We have just explained about rounding . This is what we do whenever we wish to keep reduced precision rather permanently -- when we expect all later calculations to be based on the numbers that have been rounded.

When we wish to reduce precision only temporarily, which we will occasionally wish to do -- for instance:

- in order to make a rough verbal description.
- in order to summarize which values are in a list.

we will cut the numbers to lesser precision by merely dropping one or more digits on the right.

* growing *

Sometimes we want to simplify numbers while making them somewhat bigger. We attach the word grow to these simple rules:

- numbers that are not integers grow to the next larger integer. Example: 8.7 grows to 9.
- numbers that are integers grow halfway to the next larger integer. Example: 6 grows to 6.5.

* units and decimal points *

Numbers can be written down for storage -- live or dead -- or examination. We will almost always write them down to be looked at, either as part of the description of a computation or as part of results -- final or intermediate.

When writing down a single number, keeping both raw units and decimal points often helps. To say that there are 200,000,000 million people in the U.S., or that a sheet of plastic film is 0.00127 inches thick is likely to convey about as much as we can.

Once we have a few numbers to look at, it usually helps us to move to easily understandable units that are more helpfully related to the numbers at hand. To give a population as 201 millions -- or 201.2 -- is likely to be seen more clearly than as 201,234,567. To give the thickness of a plastic film as 1.27 thousandths of an inch, instead of 0.00127 inches, is equally likely to be helpful.

Once we have more than a few numbers to look at, decimal points are a liability. They break up the appearance of the numbers, making it harder to make comparisons, especially among more than two numbers. The cure is to shift to units that are as convenient for the numbers at hand as possible. Instead of 201.2, 127.3, 63.4,...millions of people, we do better to work with 2012, 1273, 634,...hundreds of thousands. Instead of

thicknesses of 1.27, 2.52, 0.62, 3.83... thousandths of an inch, we do better to work with 127, 252, 62, 384,...hundreds of thousandths of an inch.

To be clear about what we are doing, it is often convenient to label parts (panels) of table with the units involved. Thus such headings as:

A) POPULATION -- in hundreds of thousands

A) POPULATION -- unit -- 100,000 people

A) POPULATION -- unit = 10^5 people

B) FILM THICKNESS -- unit = 0.00001 inch

B) FILM THICKNESS -- unit = 10^{-5} inch

will occur fairly often.

* h for "and a half" *

Even if the data is in whole numbers (integers), the numbers we calculate from it are often fractions. We try to keep our arithmetic simple. As a result quite a few of these fractions are halves.

We would like to avoid 4.5, 13.5 and other distracting appearances of ".5". While not yet used extensively, the use of an "h" without space or decimal point to stand for ".5" or "and a half" is very helpful. We will often use it (especially in later chapters).

When we want to count by halves, then we may go 4, 4h, 5, 5h, 6 instead of 4, 4.5, 5, 5.5, 6 or 4.0, 4.5, 5.0, 5.5, 6.0.

* # *

We shall use often before "of", "#", for "count" or "number" wherever its use seems helpful.

1C. Scratching down numbers

Probably the simplest task that the reader is likely not to be either equipped or accustomed to do is to write down a batch of numbers in such a way as to give a general feel of "what they are like". Here, by a

batch

of numbers we mean a set of similar values, obtained however they may have been. Simple examples might be (i) the weights of 21 students in a freshman class, (ii) the total winter snowfall at a selected place for each of the last 10 winters, (iii) the total sales this year of the 14 life insurance agents who sold the most last year, (iv) the number of electricity failures during the last decade in each of 11 high-income suburbs of New York City, (v) the numbers of mites found on each of 49 rats.

In each of these cases we are likely to desire a feel for the set of values as a whole. We may also want to copy down from whatever source we may have found them in. If one technique will serve both ends, so much the better. This short chapter will be devoted to learning a few ways in which we can do just this.

Let us begin with the used-car advertisements in the Sunday Standard-Times of New Bedford, Massachusetts for 18 August 1968. Some 18 ads list three or more cars. The numbers listed are, in order: 6, 7, 7, 3, 5, 7, 3, 11, 16 4, 17, 17, 6, 7, 9, 4, 7, 5. Just writing them down gives some idea of how many cars to an ad. Surely, however, we can do better. Exhibit 1 shows a way to write these numbers down that gives us more of a picture of them as a whole. In panel A each number is indicated by an "X". In panel B we gain a little by substituting the last digit for the X's, thus giving us something more to grasp with the flying eye.

exhibit 1 of chapter 1

Tallying numbers of cars per advertisement in such a way as to give some feeling for the 18 numbers AS A WHOLE

A) CONVENTIONAL

3	XX
4	XX
5	XX
6	XX
7	XXXXX
8	
9	X
10	
11	X
12	
13	
14	
15	
16	X
17	XX

B) DIGIT-REMINDER

3	33
4	44
5	55
6	66
7	77777
8	
9	9
10	
11	1
12	
13	
14	
15	
16	6
17	77

1D. Doing better with stem-and-leaf

The technique we have just been learning works well in some cases, but it is easy to find cases where it is not good enough. If we turn to the prices of second-hand Chevrolets (including Chevelles and Impalas) in the same 18 ads, we find the following (in dollars): 250,150,795,895,695, 1699,1499,1099,1693,1166,688,1333,895,1775,895,1895,795. Running the eye over this string of numbers doesn't help too much. Neither would a direct application of the technique of the last section, since going from 150 to 1895 in single dollar steps takes 1746 lines, of which only 17 would be marked.

We could, of course, drop off the last two digits, and keep only the complete hundreds of dollars. Cutting the numbers in this way would give: 2,1,7,8,6,16,14,10,16,11,6, 13,8,17,8,18,7, which could easily be treated as in exhibit 1. We shall set this possibility aside for the moment, planning to come back to it shortly.

Let us return to panel B of exhibit 1. Exhibit 2 sets out the data when each line corresponds to two adjacent number of cars per ad -- and when each corresponds to three adjacent numbers. Panels C and D of exhibit 2 show what happens when we condense 10 possible values on each line.

The panel-D version is quite efficient. One has even less to write down than if one were to write the given values down as is. Consider, for example, 1|167, which uses five characters to represent 11,16,17. This as-is form requires six characters -- or nine if we count punctuation -- rather than five.

exhibit 2 of chapter 1

The data and notation of panel B of exhibit 1 applied when each line includes 2, 3 or 10 possibilities

A) 2 per LINE

2-3	33
4-5	4455
6-7	6677777
8-9	9
10-11	1
12-13	
14-15	
16-17	67

B) 3 per LINE

3-5	334455
6-8	6677777
9-11	91
12-14	
15-17	67

C) 10 per LINE

0-9	33445566777779
10-19	167

Note: Entries are final digits of observed numbers.

D) same as PANEL C, ABBREVIATED -- to be called STEM-and-LEAF

0	3344556677779
1	167

The central idea of panel D is to give part of the information once for all at the start of each line and then give the rest of the information in the line, packed together as closely as makes sense. We shall call such displays

stem-and-leaf displays

Each line is a s t e m , each piece of information on a stem is a l e a f . When, as here, the label for the stem is the first part of a number, to be followed by each leaf in turn, we call the label the s t a r t i n g p a r t .

We like to use one-digit leaves, but -- as we will see shortly -- we can find the use of two- or more-digit leaves a good thing.

Any stem-and-leaf display is in terms of some unit. For our used car prices, for example, $10 seems a convenient unit. Thus $250 becomes 25 and appears as 2|5.

What shall we do with $1099? We could cut to 109 or round to 110. Which should we do? The general principle with which the introduction to this chapter closed is simple. If all further calculations are to be based directly on the values in the stem-and-leaf display, we should round. If not, we should cut.

Panel C of exhibit 3 shows the result of cutting our Chevrolet prices to a $10 unit and giving the results in standard stem-and-leaf form, namely with:

- all but the last digit remaining as the starting part, labelling the stem,

exhibit 3 of chapter 1

Three displays of the chevrolet prices

A) The PRICES -- in dollars
250,150,795,895,695,1699,1499,1099,1693,1166,688,1333,895,
1895,795

B) UNIT = $100
CHECK-OFF

1	X
2	X
3	
4	
5	
6	XX
7	XX
8	XXX
9	
10	X
11	X
12	
13	X
14	X
15	
16	XX
17	X
18	X

C) UNIT = $10*
STEM-AND-LEAF
one-digit leaf

1	5
2	5
3	
4	
5	
6	98
7	99
8	999
9	
10	9
11	6
12	
13	3
14	9
15	
16	99
17	7
18	9

D) UNIT = $1
STEM-AND-LEAF
two-digit leaf

1	50,
2	50,
3	
4	
5	
6	95,88,
7	95,95,
8	95,95,95,
9	
10	99,
11	66,
12	
13	33,
14	99,
15	
16	99,93,
17	75,
18	95,

* Values cut, not rounded.

- the last remaining digit entered on that line as the leaf.

This display also shows the prices in $100 units á la panel A of exhibit 1 and the full prices in a different stem-and-leaf display with two-digit leaves. Clearly:

- each of the three panels conveys a quite similar general picture of the 17 prices as a whole.

- the amount of detailed information increases as we go from panel B to panel C to panel D.

- the cost of this increase in information is negligible in going from panel B to panel C, but may not be negligible in going from panel C to panel D.

As a result, we see that we will do well to use the style of either panel C or panel D in such a situation as this.

* checking *

One thing we regretfully learn about work with numbers is the need for checking. Late-caught errors make for painful repetition of steps we thought finished.

Checking is inevitable, yet if it is too extensive we spend all our time getting the errors out of the checks. Our need is for enough checks but not too many.

In scratching down data in a stem-and-leaf, we ought to make at least one check, to ask one question: Have we made as many entries as there were values to enter? If there are but few numbers in all, we can point at them in turn and count them up without writing anything down. For the present, though, in order to remind us of the need for checking, we will

carry a "(check count)" column in parentheses. (An example will appear in the next exhibit.) To keep the calculation of such check columns from being an effort, we need to learn how to count the number of leaves on a stem rapidly and easily. Consider the 19 leaves that will appear on one stem of the next exhibit, namely

0121243122301214202

The recommended procedure is the following:

- place the left hand near the paper, with the fore-finger pointing to the right.
- cover blocks of three digits at a time counting "3", "6", "9",... to one's self as this is done.

In this example, we would have, successively -- the black figures are visible, the light ones concealed:

0121243122301214202	"0"
0121243122301214202	"3"
0121243122301214202	"6"
0121243122301214202	"9"
0121243122301214202	"12"
0121243122301214202	"15"
0121243122301214202	"18"

Clearly "18" and "1" make 19.

If we find an error in the total count, or wish to check more carefully, we need to see whether we have the right leaves on the right stems. The easy way to do this is to go through the given values again, placing a dot over each leaf

that is (either originally or after adjustment) correct. This process goes rapidly and seems quite effective.

1E. Using the right number of stems

Sometimes splitting stem from leaf between adjacent digits does not seem satisfactory. One split involves so few stems that the values are too crowded for us to see them at all well, yet moving one more digit into the starting part makes so many stems that we are bothered in the opposite way.

* expanded stem-and-leaf *

One way out is to use two stems -- two lines -- for each starting part. In such a display, we use one line for leaves 0,1,2,3,4 and the other for leaves 5,6,7,8,9. The starting parts are often best repeated.

Exhibit 4 shows an example using the land areas of the 82 counties of Mississippi. In this instance, using the conventional stem-and-leaf with one-digit starting parts would have not have worked badly (as the reader may easily verify). However, the two-line version gives us a better picture of how county areas behave in Mississippi.

* compressed stem-and-leaf *

Sometimes we wish to move in the opposite direction, squeezing two lines together. Here the device of separating both the label and the leaves by a colon works fairly well. Exhibit 5 shows an example from a later chapter. Clearly this technique could be extended to three or more starting parts per line if and when appropriate.

exhibit 4 of chapter 1

The areas of the counties of Mississippi
in a two-line display

A) SAMPLE DATA - in square miles

448,405,729,724,412,917,592,...

B) Same ROUNDED - to 10's of square miles

45,40,73,72,41,92,59,...

C) The DISPLAY - in 10's of square miles

(check) (count)		
(1)	3	8
(19)	4	0121243122301214202
(12)	4	597886556569
(6)	5	142010
(12)	5	977899958797
(6)	6	412441
(6)	6	898598
(9)	7	320341203
(5)	7	86657
(3)	8	303
(1)	8	8
(2)	9	24
(82, check)		

D) NOTES

Doubtful roundings (originally ending in 5) are made to nearest **even** number.

Source: The World Almanac, 1966, page 370.

exhibit 5 of chapter 1

Compressing a straggly stem-and-leaf

A) ORIGINAL STEM-and-LEAF

(1)	11	8
(1)	12	0
(3)	13	488
(2)	14	08
(4)	15	1266
(3)	16	058
(2)	17	08
(2)	18	58
(·)	19	
(2)	20	03688
(2)	21	38
(1)	22	1
(1)	23	5
(2)	24	05
(1)	25	8
(1)	26	3
·	27	
·	28	
(2)	29	5
·	30	
(1)	31	2
·	32	
(1)	33	0
·	34	

B) COMPRESSED VERSION

(1+1 = 2)	11:12	8:0
(3+2 = 5)	13:14	488:08
(4+3 = 7)	15:16	1266:058
(2+2 = 4)	17:18	08:58
(0+5 = 5)	19:20	:03688
(2+1 = 3)	21:22	38:1
(1+2 = 3)	23:24	5:05
(1+1 = 2)	25:26	8:3
(0+0 = 0)	27:28	:
(2+0 = 2)	29:30	35:
(1+0 = 1)	31:32	2:
(1+0 = 1)	33:34	0:

C) NOTES

Source: Example in later chapter, related to exhibit 6, following.

* using mixed leaves *

Sometimes we use stem-and-leaf more as a way to collect numbers than as a way to look carefully at them. When we are doing this we may be able -- or we may need -- to do something special with values that trail out away from the bulk of the batch.

Two situations deserve special attention:

- a few values trailing out to one or both sides of a central clump.

- many quite small values grading into a few quite -- or very -- large ones.

A useful approach to the second situation is illustrated in exhibit **6**, where our leaves jump from a single digit to two as we cross 100 -- and from two to three as we cross 1000. Notice, in addition to the check count:

- the strong horizontal bars indicating each change in technique, carried all the way across.

- the use of additional vertical bars to denote digits transferred from stem to leaf (by comparison with the top of the display). Thus, in this display, 99 and 101 would be represented as

9|9 and **1**||01,

respectively.

- the use of corners joining (left) horizontal bars to (outward) vertical bars to stress further each change in number of digits.

exhibit 6 of chapter 1

Ultimate power of hydroelectric plants of the Bureau of Reclamation (in megawatts)

A) SAMPLE VALUES

30, 1345, 225, 900, 120, 162, ...

B) STEM-and-LEAF -- with mixed leaves

```
(2)     1 | 65
(4)     2 | 5244
(5)     3 | 03626
(3)     4 | 508
(1)     5 | 0
(1)     6 | 0
(2)     7 | 51
        8 |
        9 |
----------+-------------------------------
(8)     1 | 20,62,34,50,00,20,14,08
(3)     2 | 25,85,50,
(1)     3 | 79,
(1)     4 | 24,
        5 |
        6 |
        7 |
        8 |
(1)     9 | 00,
----------+-------------------------------
(2)     1 | 345,974,
        2 |
(34, check)
```

C) NOTE

Source: The World Almanac, 1966, p. 263 (1967, p. 267).
Their source: U. S. Bureau of Reclamation.

This type of stem-and-leaf is very effective in accumulating information about values spread from many quite small to a few quite large. It does not, however, leave us with a clear feeling about the numbers. Since the numbers tend to pile up on the "1" of each new digit -- an example of what is sometimes called the "abnormal law of large numbers" (Benford?, 19--) -- there tends to be a triangular appearance for each "1" to "9" -- as if the leaves had washed against the dividing lines -- an appearance which covers up most of whatever interesting things we might see. (We will learn how to deal with this before long.)

We can often deal with the first situation in an entirely similar way, using larger leaves for the trailing ends.

It often makes things even simpler, however, to feel free to jump to longer leaves in the middle of a decade, as illustrated in exhibit 7 where one-digit leaves are used from -59 to +59, and two-digit leaves begin at -60 and +60. This enables us to use 16 lines where 31 (+11 to +0 and -0 to -18) would have been required if all leaves had been single digits.

Notice in this exhibit also, in addition to check counts and strong horizontal bars:

- the use of stem-and-leaf for both positive and negative values -- and the consequent need for both a "+0" stem and a "-0" stem.

exhibit 7 of chapter 1

Mixed-leaf stem-and-leaf applied to 25 values,
some trailing

A) The VALUES

5, -52, -27, -83, 8, -14, -122, -110, 112, 58, -119, 33, 18,
-52, -19, 12, -82, 14, 25, -182, -40, 64, -56, 5, 13.

B) The STEM-and-LEAF

(1)	1	12
(1)	0	64
(1)	5	8
	4	
(1)	3	3
(1)	2	5
(4)	1	2438
(3)	0	585
	-0	
(2)	-1	49
(1)	-2	7
	-3	
(1)	-4	0
(3)	-5	262
(2)	-0	83,82,
(4)	-1	22,10,19,82

(25, check)

1F. How to count by tallying

The basis of stem-and-leaf technique, entering an additional digit -- or digits -- to mark each value, works well for batches of limited size. Once we have much more than 20 leaves on a stem, however, we are likely to feel cramped -- and our stems begin to be hard to count. We ought to be able to escape to some other way of handling such information.

The fast methods involve one pencil (or pen) stroke per item. One method counts by fives in this style:

/ // /// //// ~~////~~

This has been widely used. The writer finds it treacherous, especially when he tries to go fast. (It is too easy for him to do

~~///~~ or ~~/////~~

for satisfactory performance.)

The recommended scheme uses first dots, then box lines, then crossed lines to make a final character for 10. Thus:

4 is

8 is

10 is

The order in which the four dots are placed around the square is unimportant, as is the order in which the four sides of the square are filled in, and the order in which the two diagonals are crossed.

Once anyone is used to this scheme, the occurrence of errors is much less than for the other. The closure of the four dots completing the outline of the square is clear and definite, there is no tendency to go on to lines after three points -- or to add a fifth point. The same applies to the completion of the sides of the square. (A few novices report a tendency to move on to the next character before crossing the diagonals. Their eights -- ☐ -- are, however, distinctive enough from tens -- ⊠ -- to be picked up once the tallying has been completed and the result is being converted into digits.

Exhibit 8 shows a variety of simple examples.

(* used cars again *)

For exhibit 9 we return to the Sunday Standard-Times and its used-car ads. This display gives -- in both tallied and digit form -- the number of cars of each model year advertised:

- in ads for one or two cars.
- in ads for three or more cars.

We can get a quick grasp of appearances easier and faster from the tallies-by-ten than from the same counts in digits. (It may even be worthwhile, on occasion, to turn digit-given counts into tallies-by-ten to look at them.)

exhibit 8 of chapter 1

Tallying by tens variously illustrated

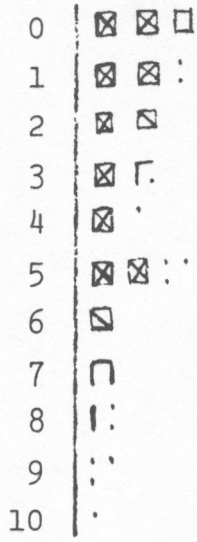
A) NUMBERS -- in tally and digits

Tally	Number
.	1
. . or . .	2
. . . or . . (etc.)	3
. . . .	4
—. . or | . . (etc.)	5
⌐. or ⌊. (etc.)	6
⊓ or ⊏ (etc.)	7
□	8
⧅ or ⧄	9
⊠	10
⊠ ⊠ ⊓	27
⊠ ⊠ ⊠ ⊠ . .	42

(Cont'd)

exhibit 8 (Cont'd)

B) A SIMPLE EXAMPLE



C) A TWO-WAY EXAMPLE

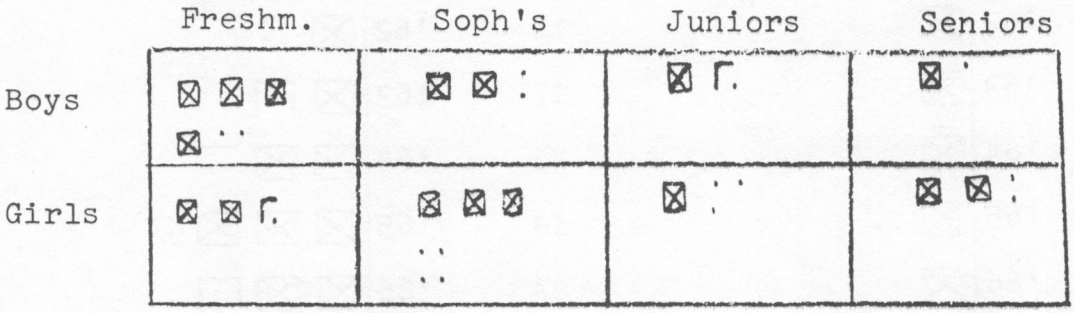


exhibit 9 of chapter 1

Model-years of used cars advertised

A) ADS for 1 or 2 CARS

Digits	Tally	
1	'54	·
2	'55	··
-	'56	
1	'57	·
2	'58	··
5	'59	‾‾ ··
5	'60	‾‾ ··
2	'61	··
11	'62	⊠ ·
12	'63	⊠ ··
10	'64	⊠
14	'65	⊠ ·· ··
12	'66	⊠ ··
13	'67	⊠ ··
3	'68	··

B) ADS for 3 or MORE CARS

Digits	Tally	
-	'54	
1	'55	·
-	'56	
1	'57	·
1	'58	·
1	'59	·
4	'60	··
10	'61	⊠
14	'62	⊠ ··
17	'63	⊠ ⊓
25	'64	⊠ ⊠ ‾‾ ··
34	'65	⊠ ⊠ ⊠ ··
27	'66	⊠ ⊠ ⊓
12	'67	⊠ ··
4	'68	··

C) NOTE

Source: Sunday Standard-Times, New Bedford, Mass. for 18 August 1968 (page 51, Fourth Section).

The general impressions easily gained from this display include:

- a tendency for small ads (one or two cars) to offer somewhat older cars.
- a question as to why '55 appears to be popular.

Each of us can provide or invent reasons for the first of these. As for the second, reexamination of the ads shows that two of the three 55's are "classic Thunderbirds".

While tally-by-tens is an effective technique, it does not tell us as much as stem-and-leaf. It should be almost exclusively used as an overflow procedure. Indeed, overflow need not occur as often as we might think, for it is often very helpful to accumulate information in parts -- parts which we are then free to combine or keep apart. Display 9 shows collection in two parts -- had we gone further and collected in three or four parts, it would have been reasonable -- so far as size of counts go -- to collect each part using stem-and-leaf. If there were further information to collect, it would be not only reasonable but desirable to do this.

1G. <u>What does it mean to feel "what the data are like"?</u>

We have been looking at well-behaved batches of numbers. True there was something to pick up about 1955 used cars, but generally we saw little striking in any of our stem-and-leaf displays.

Exhibit 10 shows the height of the highest point in each of the 50 states. Clearly the states fall into three groups which turn out to be:

- Alaska.
- The Rocky Mountain States and Hawaii.
- all others.

The things that we most frequently notice in stem-and-leaf displays are:

- appearances of separation into groups.
- unsymmetric trailing off, going further in one direction than another.
- appearance of unexpectedly popular (or unpopular) values.
- about where the values are "centered".
- about how widely the values spread.

The stem-and-leaf itself guides us to the first three. A little help with the last two may be useful; in fact, we shall shortly spend a chapter or two on them.

exhibit 10 of chapter 1

The eights of the highest points in each state

A) STEM-AND-LEAF -- unit 100 feet

(5)	0	43588
(6)	1	237886
(6)	2	484030
(5)	3	45526
(5)	4	80149
(5)	5	34307
(3)	6	376
(1)	7	2
(1)	8	8
	9	
	10	
(1)	11	2
(3)	12	768
(5)	13	81258
(3)	14	544
	15	
	16	
	17	
	18	
	19	
(1)	20	3

(50, check)

B) NOTES

Source: **The World Almanac, 1966,** page 269.
Their source: National Geographic Society.

*****

1H. How far have we come?

This first chapter has been devoted to (i) a discussion of what exploratory data analysis is all about, and (ii) an introduction to the simple tools that allow us to scratch down information about batches of numbers in ways that:

- are at least as fast and easy as writing the numbers down separately (and often faster and easier)
- give us a feeling of how each batch seems to behave as a whole.

* * *

Exploratory data analysis is detective work -- in the purest sense -- finding and revealing the clues.

(In real life, detectives often have to go further, and try to assess the strength of the evidence given by the clues, as when they need to forecast the chances of a conviction. Evaluation is finally a matter for judge or jury, however. The corresponding questions about data are matters for confirmatory data analysis.)

As this chapter has begun to illustrate, one thing the data analyst has to learn is how to expose himself to what his data is willing -- or even anxious -- to tell him. Finding clues requires looking -- in the right places and with the right magnifying glass.

Most of us know womething about looking at a single number. It may not be easy to draw all the implications from the fact that there were 46,930 deaths from motor vehicle accidents in the

U.S.A. in 1964, but the difficulty is not in understanding the one number.

The same is often true of comparing a few firmly-based numbers, like these 1962-64 average death rates per hundred million person miles:

- trains	0.09
- planes	0.16
- buses	0.17
- cars (turnpike)	1.2
- cars (general)	2.3

Again we are likely to have difficulty with the implications, but not with understanding the numbers themselves.

The simplest case where we need to learn to grasp the numbers more firmly or more rapidly comes when we face a batch of similar numbers. Hence the techniques of this chapter and those that follow.

The main techniques we have met are all closely related. There is no substitute for displaying the numbers before our eyes. Putting down a standard mark for each occurrence does this. (Panels A and B of exhibit 1 and panel B of exhibit 2 illustrate the simple case.)

Putting down a mark that tells us something more about the particular value is:

- equally fast.
- easier to check.
- more informative.
- for many, more fun.

Accordingly, we plan to use stem-and-leaf techniques wherever they seem to fit. (And to at least check the total count.)

We have a choice among one-digit leaves, two-digit leaves, etc., and mixed leaves. We make this partly for convenience, partly in view of our purposes, partly by necessity. In certain kinds of situations -- many values quite small, a few quite large is the standard example -- stem-and-leaf does well in collecting numbers but not well in showing us their behavior. (We have more to learn.)

When we have many values to collect or display, and can give up further details, tallying-by-tens offers us real advantages.

(The end of the section called "How far have we come?" is the formal close of this chapter -- and of each chapter that follows. Any further sections are -- in one sense or another -- optional. Some of them may introduce convenient techniques, but these techniques are not essential to our main aims or practices.)

*****

1K. <u>How to use stem-and-leaf to pick up additional information</u> (optional technique)

The basic principle -- stem to locate where we write it down; leaf to say more -- need not be confined to the sort of examples we have so far seen. Let us return to the Sunday Standard-Times again, and the 18 ads for three or more cars. Let us now use the stem-and-leaf technique to collect combined information on the model-year and name of the cars offered. Exhibit 11 shows the technique and the results. Some aspects of the coding scheme are worth noting:

- we have used suggestive letters where available, but have not hesitated to step out to numbers when convenient letters are not available.

- we have distinguished capital and small letters whenever their handwritten forms are clearly distinct, but have not tried to use both otherwise.

Panel C of this exhibit shows the result of organizing the entries alphabetically for each model-year (and providing two lines for '65). This makes quite clear a tendency for certain name-year combinations to appear frequently. The most obvious example is the 15 '65 Chevrolets, but others can be found.

exhibit 11 of chapter 1

Collection of model-year and name information
for advertised used cars

A) The NAME CODE

Bonneville	b
Buick	B
Cadillac	A
Chevelle	h
Chevy II	2
Chevrolet	C
Chrysler	H
Classic	l
Comet	3
Corvair	4
Dodge	D
Eldorado	E
Ford	F
Futura	f
Galaxie	g
Grand Prix	G
Impala	I
Jaguar	J
Lincoln	L
Mercury	M
Mustang	m
Oldsmobile	O
Plymouth	p
Pontiac	P
Rambler	r
Simca	5
Studebaker	s
Sunbeam	S
Tempest	t
T-bird	T
Valiant	V
Volkswagon	v

B) The COLLECTION

'55	T
'56	
'57	E
'58	C
'59	I
'60	AIAv
'61	CItHvFFFHT
'62	4CIP203CTIIP2
'63	IIFrBFrCrLOICCbP
'64	PglBDD4CIIrBI3PshCAOCM
'65	5H4C40CpDCFCCCbBICSCPOIDCCBACC2CCCr
'66	ftvhlCv3B4hFIvACOhmPJFFCp3
'67	VPPFGmvPmCAG
'68	HpDp

C) The COLLECTION ORGANIZED

'55	T
'56	
'57	E
'58	C
'59	I
'60	AAIv
'61	CFFFHHItTv
'62	CCIIIOPPT2234
'63	bBCCCFFIIILOPrrr
'64	ABBCCCDDghIIIMOPPrs134
'65	CCCCCCCCCCCCCCC
'65	AbBBDDFHIIOOpPrSW245
'66	ABCCCfFFFhhhIJmOpPtvuvl334
'67	ACFGGmmPPPvV
'68	DHpp

D) NOTES

Source: Ads for three or more cars, page 51, Fourth Section, Sunday Standard Times, New Bedford, Mass. for 18 August 1968.

******

1P. Problems for Chapter 1.

1. In 1963, McGlanery and Harbein gave the values in exhibit 12 showing how well they could measure the viscosity of liquids with a device called a capillary vreometer. Make appropriate stem-and-leaf displays for each of the three samples, comment on the appearance of each.

2. In 1958, Heckmann and Lübeck gave the radial velocities in exhibit 13 for stars near the star Alpha Persei. Make an appropriate stem-and-leaf comment on its appearance.

3. In 19 , Hindman gave the numbers in exhibit 14 which refer to the stability of a complex ion combining 4-valent Plutonium ions with nitrate ions. Make an appropriate stem-and-leaf diagram. Comment on its appearance.

exhibit 12 of chapter 1

Viscosity at a temperature

of 190°C and a low rate of shear

Run (for) each sample	Viscosity in 100,000's of poises		
	Sample I	Sample II	Sample III
1	.384	.661	3.54
2	.376	.671	3.66
3	.376	.688	3.42
4	.371	.644	4.10
5	.385	.668	4.09
6	.377	.648	3.77
7	.365	.706	4.17
8	.384	.715	3.91
9	.365	.647	4.61
10	.384	.682	3.87
11	.373	.692	
12		.729	

Source: R. M. McGlanery and A. A. Harban 1963. Two instrumetns for measuring the low-shear viscosity of polymer melts. Materials Research and Standards 3: 1003-1007. Table 2 on page 1004.

exhibit 13 of chapter 1

Radial velocities of stars near α Persei

A) The RADIAL VELOCITIES -- in kilometers per second

+50.0	-2.4	+1.0	-0.5	+2.2	-11.8	-11.9
-36.1	+10.0	+3.7	-15.9	-4.1	-19.1	-16.2
-7.0	+7.0	-9.5	+2.0	-3.0	-22.1	-10.8
+2.2	0.0	-8.4	+5.0	+3.2	-9.1	-6.0
-2.8	-1.0	+0.5	+4.4	-4.4	-17.2	-23.9
+3.0	-0.7	+2.2	+1.6	-0.3	+12.8	-8.2
+24.7	+15.9	+18.0	+6.0	+14.5	-10.5	-13.6
+4.8	+9.0	-17.5	+5.9	-18.4	-17.2	-4.4
						+15.9
						-25.7

B) The SOURCE

O. Heckmann and K. Lübeck 1958. Das Farben-Helligkeits-Diagramon des Bewegungshaufens um Alpha Persei. Zeitschrift für Astrophysik 45: 243-263. Figure 2 on page 248 is based on tables 2 and 3 on pages 247 and 248.

exhibit 14 of chapter 1

Association constant of the first nitrate complex ion of Plutonium (IV) measured in perchloric-nitric acid solutions with 2.00 M hydrogen ion concentration

A) RESULTS of REPEATED EXPERIMENTS

1.97, 2.13, 2.02, 1.55, 2.83, 2.92, 2.40, 3.24, 3.21, 3.88, 3.85, 3.89, 2.84, 3.08, 4.17, 2.64, 3.67, 3.06 2.70, 1.91, 2.34, 2.49, 2.76, 3.50

B) The SOURCE

J. C. Hindman 19 . Complex ions of plutonium. The nitrate complex ions of plutonium (IV). Pages 388-404 of The Tranurani Elements; Edited by Table 1 on page 390.